2020 Census Disclosure Avoidance System

DEVELOPMENT & RELEASE TIMELINE

We're modernizing our approach to privacy protection for the 2020 Census in the face of new, digital-age threats.

A powerful new disclosure avoidance system (DAS) designed to withstand modern re-identification threats will protect 2020 Census data products (other than the apportionment data; those state-level totals remain unaltered by statistical noise).

Inspired by cryptographic principles, the 2020 DAS is the only solution that can respond to this threat while maximizing the availability and utility of published census data.

1981

The U.S. Census Bureau (USCB) forms the Microdata Review Panel (MRP) to ensure microdata files aren't released that allow re-identification. Microdata files contain non-aggregated records for individual responses, unlike aggregated statistics published in data tables.

1995

Growing concerns about the adequacy of disclosure avoidance protections for all data products, not just microdata files, leads the USCB to expand the scope and recharter the MRP as the Disclosure Review Board (DRB). The change was driven by two primary forces: new, post-1990 Census plans to release data that was more granular than in previous censuses, and the need to standardize disclosure avoidance processes for the 1992 Economic Census.

2000

The efficacy of existing privacy protections is called into question when an external researcher finds that 87% of the population in the 1990 Census (216 million people) are unique on just three pieces of information: zip code, gender, and date of birth.[1]

2001

High-profile re-identifications of people in public databases prompts the USCB to form the Data Stewardship Executive Policy Committee (DSEP), a permanent, executive-level body overseeing decision-making and communications on policy issues related to privacy, security, confidentiality, and administrative records. The new DSEP oversees the work of the long-standing Disclosure Review Board (DRB).

2003

Data scientists Irit Dinur and Kobbi Nissim prove that if too many statistics are published too accurately the contents of a confidential database will be exposed with near certainty after a finite number of queries. A necessary condition to avoid this outcome is infusing noise into every released statistic.[2]

2006

Data scientists Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith create new algorithm-based, cryptography-inspired anti-disclosure protection method to protect against digital-age threats.[3]

2008

USCB's *OnTheMap* tool becomes world's first production application of differential privacy.[4]

2010

The Census Bureau's Disclosure Review Board expresses concerns about releasing block-level data, given the decline in block-level populations and the growth in publicly available corroborating data sources. The DSEP concurred that these issues need to be factored in to 2020 Census planning.

2012

USCB and external researchers conclude that targeted re-identification is "fairly straightforward" given the large amount of public data available. The study attempted to link the American Community Survey (ACS) Public Use Microdata Sample (PUMS) with a publicly available external dataset. It tentatively identified 389 of the 2.3 million people in the dataset (0.017%), of which 87 were correct. Given these rates, the study concludes that large scale re-identification is less likely as of 2012.[5]

2014

An external researcher's claim that he was able to re-identify a housing unit record in the Census Bureau's New York City Housing and Vacancy Survey prompts an internal re-identification study. The study concludes the risk for this survey is small, but identifies vulnerable variables, which the Census Bureau subsequently removed or recoded.

An internal analysis of the growing potential of published data re-identification prompts the DSEP to commit to prioritizing research into new disclosure avoidance methods.

2016

Jun

Chief Scientist briefs DSEP on growing threats to the Census Bureau's ability to protect respondent confidentiality using legacy systems. DSEP directs Chief Scientist to develop enhanced privacy protection methods.

Dec

Chief Scientist forms a team to start researching the extent to which 2010 Census disclosure avoidance methods can withstand current reconstruction and re-identification attacks. The study uses a three-step process:

(1) Attempt to first reconstruct records using only P.L. 94-171 Redistricting Data Summary File and Summary File 1 published data.
(2) Attempt to link the reconstructed microdata to 2010-era commercial name and address data.
(3) Compare the results to the confidential 2010 Census Edited file (CEF).

2017

Apr Early results of the re-identification study, focused on a single county, reveal a 100% reconstruction rate. Per risk management protocol, this discovery shifts the threat status, triggering the requirement for a mitigation plan for the 2020 Census. Researchers expand the focus to attempt an attack on the entire U.S.

Sept (9/7) The Commission on Evidence-Based Policymaking recommends that the President direct Federal departments to adopt state-of-the-art database, cryptography, privacy-preserving, and privacy-enhancing technologies for confidential data used for evidence building.

2018

Feb (2/15) DSEP approves steps to mitigate disclosure risks across Census Bureau data products. Measures include limiting published 2020 Census data (other than the required P.L. 94-171 redistricting files) to those with data user-identified use cases.

The Census Bureau released billions of statistics from the 2010 Census redistricting data alone. Data prioritization is needed because the more statistics we publish, the easier it is to reconstruct responses and re-identify people. The alternative to reducing the number of released statistics would be a scale of table suppression and noise addition that would be unacceptable.[6]

Apr (4/30) DSEP issues a moratorium on release of high-risk Census Bureau data products unless protected with enhanced legacy noise-infusion methods or differential privacy mechanisms.

Jul (7/19) USCB solicits data use cases via a Federal Register notice to help prioritize the final slate of 2020 Census published data products and the fitness-for-use needs of each.

Oct (10/09) USCB extends the comment period for data use cases.

Nov (11/8) DSEP sets invariants (variables to be published as enumerated without statistical “noise”) and the privacy-loss budget (PLB) for the prototype P.L. 94-171 (redistricting) file based on the 2018 Census Test.

2019

Feb (2/16) USCB announces completion of the re-identification research begun in 2016. The research demonstrated that modern tools were in fact able to exactly reconstruct the block location and voting age – without names or addresses – of 100% of the U.S. population in the 2010 Census. About 71% of the population in the confidential census edited file, or “CEF,” was reconstructed when race, Hispanic/Latino ethnicity, sex and age (+/- one year) data were added.

Researchers were able to connect names and addresses to the reconstructed records for almost half of the population using 2010-era commercial data. About 52 million people enumerated in 2010 (17%) were correctly re-identified. This is four orders of magnitude greater than the research on the ACS PUMS data from 2012, which reinforced the need for significantly stronger privacy protections for 2020 Census data.

Mar (3/8) DSEP approves the proposed suite of 2020 Census data products.

(3/28) USCB releases the source code behind the prototype redistricting data based on 2018 Census Test data.

Apr (4/15) USCB releases the prototype P.L. 94-171 (redistricting) file based on 2018 Census Test data using an early test version of the 2020 DAS.

Jul (7/30) USCB officially announces that a differential privacy-based disclosure avoidance system will not be applied to the American Community Survey until 2025, at the earliest.

Sept (9/13) USCB presents the proposed list of 2020 Census data products based on user feedback to the Census Scientific Advisory Committee.

Oct (10/29) USCB releases first (baseline) 2010 Demonstration Data. This is the first of several releases that apply the latest iteration of the DAS to 2010 Census data for comparison purposes.

The DSEP-chosen privacy-loss budget (PLB) of 6.0 (4.0 for person records, 2.0 for housing) will be used on all development releases going forward to allow an “apples-to-apples” analysis of incremental DAS development progress. The DSEP will choose the final PLB after careful analysis, prior to final data processing.

Dec (12/11-12) Data users provide invaluable feedback on the first demonstration data at a workshop hosted by the National Academies of Science’s Committee on National Statistics (CNSTAT). Their feedback helps Census Bureau DAS developers identify additional use cases and informs ongoing development.

2020

Mar (3/27) USCB releases initial quality metrics to allow easier analysis of ongoing DAS development progress for each demonstration data release. USCB also seeks input on additional measures to help data users assess each release and allow comparisons to published 2010 data.

May (5/18) USCB launches newsletter to provide more timely DAS development updates.

(5/27) USCB releases second set of 2010 Demonstration Data and quality metrics. This produced a second round of detailed external user feedback including specific proposals for category binning and minimum-size geographic areas.

Aug (8/26) Pandemic-triggered operational delays require the Census Bureau to shift the DAS development focus to attempt to meet the pre-pandemic data deadline. Focus shifts to redistricting data product-only until further notice. This limited the Census Bureau’s ability to further develop the portions of the DAS that cover tables not included in the redistricting data product, including the ability to address many of the external suggestions from the first two demonstration data products.

Sept (9/17) USCB releases third set of 2010 Demonstration Data and quality metrics, focused on redistricting data. This data product was based on the newly slimmed down code base for the redistricting data, and it contained several implementation errors that were discovered simultaneously by several external stakeholders and the internal development team.

Nov (11/17) USCB releases fourth set of 2010 Demonstration Data and quality metrics, focused on redistricting data. This release corrected the implementation errors in the September Demonstration Data product and, once again, produced many analyses by external stakeholders that were used to refine the algorithms. In addition, this release was used to develop the final set of redistricting metrics, in collaboration with specialists in the redistricting community and at the Department of Justice.

(11/24) DSEP finalizes the list of “invariants” for the first set of 2020 Census data products. Invariants are statistics that are published without noise infusion.

Dec USCB starts experiments to evaluate different settings of key system parameters for the redistricting data. Hundreds of full-scale experimental runs of the DAS will help the DSEP determine:

- The optimal set and processing order of queries against the confidential data.
- The share of PLB allocated to the different queries.
- The share of PLB allocated to tabulations at different geographic levels (e.g., county and tract).

2021

* Planned

Mar (3/25) DSEP approves the PLB for the 5th set of Demonstration Data. The PLB is set to ensure the accuracy of racial demographics for voting districts as small as 500 individuals. It is based on redistricting use cases informed by extensive feedback from the redistricting community and the Civil Rights Division at the U.S. Department of Justice.

Apr (4/26) USCB releases 2020 Census apportionment data. Note that apportionment data are *not* affected by the DAS.

(4/28) USCB releases fifth set of 2010 Demonstration Data and quality metrics focused on redistricting data. These files used a higher PLB than previous releases that is more reflective of the anticipated final privacy/accuracy tradeoff for the 2020 Census data products. The released full privacy-loss accounting, based on the zero-Concentrated Differential Privacy (zCDP, discrete Gaussian mechanism), shows how the noise was injected. The optimized geographic spine used to improve accuracy for low-population Minor Civil Divisions and Places.

June (6/9) DSEP selects DAS parameters, including PLB, for redistricting data based on results of experimental runs and data user feedback.

Mid-June* DAS development team implements parameters set by DSEP into algorithm for application to redistricting data product.

Late June* Data production run and quality control analysis begins.

By Aug 16* USCB to provide a legacy format summary redistricting data file to all states.

Sept* USCB will release sixth and final set of 2010 Demonstration Data and quality metrics. This “production-ready” set will include the final PLB and DAS parameters that will produce the official 2020 Census redistricting data product.

By Sept 30* USCB will release the 2020 Census Redistricting Data P.L. 94-171 Summary File.

2022+ Additional Data Releases

TBD *Demographic Profiles*

TBD *Demographic and Housing Characteristics (DHC) Profiles (former “Summary File 1,” or “SF1”)*

TBD *Detailed Demographic and Housing Characteristics File (former “Summary File 2,” or “SF2”)*

TBD *American Indian and Alaska Native Summary File*

TBD *Public Use Microdata Sample (PUMS)*

TBD *Congressional District Demographic and Housing Characteristics File*

TBD *Census Briefs*

TBD *Population and Housing Tables*

To learn more, search “Disclosure Avoidance” at www.census.gov.

[1] L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

[2] Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03). Association for Computing Machinery, New York, NY, USA, 202–210. DOI:https://doi.org/10.1145/773153.773173.

[3] Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. In Halevi, S. & Rabin, T. (Eds.) Calibrating Noise to Sensitivity in Private Data Analysis Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg, 265-284, DOI: 10.1007/11681878_14.

[4] Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.

[5] Ramachandran, Aditi, Lisa Singh, Edward Porter, and Frank Nagle. 2012. Exploring Re-Identification Risks in Public Domains, Tenth Annual International Conference on Privacy, Security and Trust, IEEE, doi:10.1109/PST.2012.6297917.

[6] McKenna, Laura. 2018. Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing, Working Papers 18-47, Center for Economic Studies, U.S. Census Bureau, Handle: RePEc:cen:wpaper:18-47.